An Analysis of Bulk Data Movement Patterns in Large-scale Scientific Collaborations

2011 J. Phys.: Conf. Ser. 331 012008

(http://iopscience.iop.org/1742-6596/331/1/012008)

Journal of Physics: Conference Series **331** (2011) 012008 doi:10.1088/1742-6596/331/1/012008

An Analysis of Bulk Data Movement Patterns in Large-scale Scientific Collaborations

W. Wu, P. DeMar, A. Bobyshev
Computing Division, Fermilab, Batavia, IL 60510, USA

E-mail: {wenji, demar, bobyshev}@fnal.gov

Abstract. Large-scale research efforts such as LHC experiments, ITER, and climate modelling are built upon large, globally distributed collaborations. For reasons of scalability and agility and to make effective use of existing computing resources, data processing and analysis for these projects is based on distributed computing models. Such projects thus depend on predictable and efficient bulk data movement between collaboration sites. However, the available computing and networking resources to different collaboration sites vary greatly. Large collaboration sites (such as Fermilab, CERN) have created data centres comprising hundreds, and even thousands, of computation nodes to develop massively scaled, highly distributed cluster-computing platforms. These sites are usually well connected to outside worlds with high-speed networks with bandwidth greater than 10Gbps. On the other hand, some small collaboration sites have limited computing resources or poor networking connectivity. Therefore, the bulk data movements across collaboration sites vary greatly. Fermilab is the US-CMS Tier-1 Centre and the main data centre for a few other large-scale research collaborations. Scientific traffic (e.g., CMS) dominates the traffic volumes in both inbound and outbound directions of Fermilab off-site traffic. Fermilab has deployed a Flow-based network traffic collection and analysis system to monitor and analyze the status and patterns of bulk data movement between the Laboratory and its collaboration sites. In this paper, we discuss the current status and patterns of bulk data movement between Fermilab and its collaboration sites.

1. Introduction

Large-scale research efforts such as LHC experiments, ITER, and climate modelling are built upon large, globally distributed collaborations. For reasons of scalability and agility and to make effective use of existing computing resources, data processing and analysis for these projects is based on distributed computing models. The data produced by these projects commonly reach petabytes or tens of petabytes per year. The ability to efficiently retrieve, store, explore, analyze, and share the datasets generated by these projects is tremendously challenging. Such projects depend on predictable and efficient bulk data movement between collaboration sites. However, the available computing and networking resources at different collaboration sites vary greatly. Large collaboration sites (e.g., Fermilab) have created data centres comprising hundreds, and even thousands, of computation nodes to develop massively scaled, highly distributed cluster-computing platforms. These sites are normally well connected to outside worlds with high-bandwidth links of more than 10Gbps. On the other hand, some small collaboration sites have only limited computing resources or poor networking connectivity. Therefore, bulk data movement between collaboration sites vary greatly in terms of

performance and scale. Naturally, a question arises: ***What is the current status and pattern of bulk data movement between collaboration sites?***

The answer to this question is not readily available. This is because network traffic is difficult to monitor and analyze. Existing tools like Ping, Traceroute, OWAMP [1] and SNMP provide only coarse-grained monitoring and diagnosis data about network status [2][3]. For example, SNMP typically provides 1-minute or 5-minute average for network performance data of interest. These averages may obscure the instantaneous network status. On the other hand, packet trace analysis [4][5] involves traffic scrutiny on a per-packet basis and requires high-performance computation and large-volume storage. It faces tremendous scalability challenges in high-speed networks, especially with the current networking industry trend of migration towards 100 Gbps. Flow-based tools such as Cisco's NetFlow [6] lie in between. They produce more fine-grained data than SNMP, yet not as detailed or high-volume as required by packet trace analysis.

NetFlow [6], first implemented in Cisco routers, is the flow measurement solution in widest use today. It was first implemented as a route lookup cache for optimizing the performance of IP packet forwarding. It was later adapted to provide flow data information for statistical analysis and network accounting. Routers running NetFlow maintain a "flow cache" containing flow records that describe the traffic forwarded by the router. These flow records are exported using unreliable UDP to a computer that collects, analyzes, and archives them. For each router interface, flows are identified by important fields in the packet header. The information in these fields includes the source and destination IP address, the protocol, the source and destination port, and the type of service. If a packet does not belong to an existing flow, the router inserts a new flow record into the flow cache. Besides the fields identifying the flow, each flow record also collects other data such as the number of packets and bytes in the flow and the timestamps of the first and last packet. These records allow many kinds of analyses. To update the NetFlow cache when a packet is seen, NetFlow must look up the corresponding entry in the flow cache, create a new entry if necessary, and update that entry's counters and timestamps. Because the processor and the memory holding the flow cache cannot keep up with the packet rate in high-speed interfaces, Cisco introduced sampled NetFlow [6][7], which updates the flow cache only for sampled packets. Other vendors (e.g., Juniper, Huawei, 3COM) provide similar mechanisms for their routers. Flow-based analysis is widely used in traffic engineering [8][9], anomaly detection [10][11], traffic classification [12][13], performance analysis and security [14][15][16], etc. For example, Internet2 makes use of flow data to generate traffic summary information by breaking the data down in a number of ways, including by IP protocol, by a well-known service or application, by IP prefixes associated with "local" networks, or by the AS pairs between which the traffic was exchanged. In [17], flow data has been applied in calculating aggregated throughputs between sites. Because the level of aggregation is usually unknown to end-users, it is difficult to use this tool to detect sub-optimal data movement. In [10], the sub-space method is applied to flow traffic to detect network-wide anomalies.

Fermilab is the US-CMS Tier-1 Centre and the main data centre for a few other large-scale research collaborations. Scientific traffic (e.g., CMS) dominates the traffic volumes in both inbound and outbound directions of off-site traffic. Fermilab has deployed a Flow-based network traffic collection and analysis system to monitor and analyze the status and pattern of bulk data movement between Fermilab and its collaboration sites. We collected and analyzed the traffic flow records between November 11, 2009 and December 23, 2009.

This paper is organized as follows. In section 2, we describe Fermilab flow-based network traffic collection and analysis system, and the flow record dataset that is used in our analysis. In section 3, we discuss the status and patterns of bulk data movements between Fermilab and its collaboration sites. Section 4 concludes the paper.

2. Fermilab Flow-Based Network Traffic Collection and Analysis System

2.1. Fermilab Flow-based Network Traffic Collection and Analysis System

The Fermilab flow-based network traffic collection and analysis system is shown in Figure 1. We collect flow data from Laboratory border routers, and some internal LAN routers. For analysis, various commercial, public domain and in-house made tools were deployed over time. Flow data exported from routers are replicated into these tools as needed.

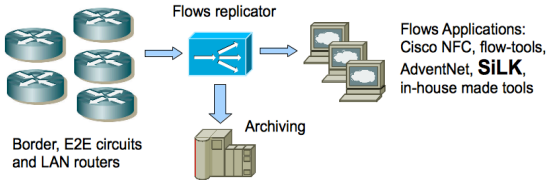


Figure 1: Flow Collection and Analysis System

Network traffic analysis described at this paper is mainly based on CMU's SiLK toolset [18]. SiLK is a collection of traffic analysis tools developed by the CERT Network Situational Awareness Team to facilitate traffic flow analysis of large networks. The SiLK tool suite supports the efficient collection, storage, and analysis of network flow data, enabling network traffic analysts to rapidly query large historical traffic data sets.

2.2. Flow Record Dataset under Analysis

We collected and analysed the traffic flow records from November 11, 2009 to December 23, 2009. The total flow record database has a size of 60GBytes, with 2,679,598,772 flow records. During this period, there were totally 23,764,073 Gbye data and 2.221x10**12 packets transferred between Fermilab and off-sites. Since Transmission Control Protocol (TCP) is the dominant protocol used for data transfer on the Internet today, we only analyzed TCP data transfers. In addition, to study the status and patterns of bulk data movements between Fermilab and its collaboration sites, we analyzed the data transfers between Fermilab and /24 off-site subnets in both inbound and outbound directions. For the sake of bulk data transmission, a /24 subnet is large enough for most collaboration sites. However, some large collaboration sties, like CERN, have multiple /24 subnets.

3. Bulk Data Movement Status and Patterns between Fermilab and its Collaboration Sites

3.1. Traffic Volume

To study the status and patterns of bulk data movements between Fermilab and its collaboration sites, we analyzed the data transfers between Fermilab and /24 off-site subnets. We counted the top 100 /24 subnets that transfer to/from Fermilab in terms of traffic volume. In the inbound direction, the traffic from top 100 off-site subnets that transfer to Fermilab amounts to 99.04% of the Laboratory's total inbound traffic. In the outbound direction, Fermilab transfers 95.69% traffic to the top 100 off-site /24 subnets. Table 1 and 2 give the traffic volume statistics of top 10 sites, both inbound and outbound. Note that some large collaboration sites might have multiple /24 subnets for bulk data transmission. For example, CERN has multiple /24 subnets that transfer to Fermilab. In table 2, subnet 128.142.178.0, 128.142.208.0, 128.142.218.0, 128.142.215.0, and 128.142.220.0 all belong to CERN. In the inbound direction, the 100 top subnets belong to 67 collaboration sites; in the outbound direction, the 100 top subnets belong to 76 collaboration sites.

Fermilab is the US-CMS Tier-1 Centre and the main data centre for a few other large-scale research collaborations. Scientific traffic (e.g., CMS) dominates the traffic volumes in both inbound and outbound directions of Fermilab off-site traffic. In the inbound direction, the aggregated traffic from

CERN amounts to almost 50% of Fermilab's total inbound traffic. In the outbound direction, the outgoing traffics are distributed across collaboration sites. There are no obvious dominating sites. Also, we noticed that the traffic between Fermilab and its collaboration sites are seriously asymmetric. This situation actually reflects the fact of the hierarchy-computing model being applied in CMS and other scientific collaborations.

Table 2. Traffic Volume Statistics of TOP 10 Sites in the Inbound direction

/24 Subnets	Bytes	%Bytes	Cumulative%
128.142.178.0	1.81261E+14	18.649967	18.649967
128.142.208.0	1.61148E+14	16.580516	35.230483
128.142.218.0	5.40668E+13	5.562923	40.793406
128.142.215.0	4.37325E+13	4.499631	45.293037
152.54.1.0	4.04937E+13	4.166388	49.459424
128.142.220.0	3.86989E+13	3.981727	53.441151
128.211.143.0	3.20479E+13	3.297406	56.738557
131.154.129.0	2.70239E+13	2.780484	59.519041
128.135.70.0	2.57182E+13	2.646146	62.165187
129.93.227.0	2.56594E+13	2.640093	64.80528

Table 3. Traffic Volume Statistics of TOP 10 Sites in the Outbound direction

/24 Subnets	Bytes	%Bytes	Cumulative%
193.48.99.0	1.04926E+14	7.817906	7.817906
129.93.239.0	7.55832E+13	5.631628	13.449535
117.103.110.0	5.49563E+13	4.094741	17.544275
130.246.179.0	4.8628E+13	3.623222	21.167498
202.122.33.0	4.67251E+13	3.48144	24.648938
169.228.131.0	4.42283E+13	3.295402	27.94434
10.31.62.0	3.83632E+13	2.8584	30.80274
193.190.247.0	3.77884E+13	2.815572	33.618312
128.211.143.0	3.60189E+13	2.683731	36.302043
129.59.197.0	3.54771E+13	2.643364	38.945406

3.2. Round Trip Time (RTT) and Circuitous Paths

The top 100 /24 subnets are scattered across the world. We collected and calculated the Round Trip Time (RTT) between Fermilab and these subnets in both inbound and outbound directions. To collect and calculate RTT, we pinged the IP addresses collected from the flow records. Multiple IP addresses were pinged for each subnet and we calculated the average. Figure 2 gives the RTT statistic for these subnets. The RTT statistics essentially indicate the physical locations of Fermilab's collaboration sites. Based on the RTT statistics, the top 100 subnets can be categorized into three groups. The first group comprise of subnets that have RTTs less than 100ms. This group of subnets represent the collaboration sites located in North America. Due to physical adjacency to Fermilab, their RTTs are less than 100ms. The second group consists of subnets that have RTTs between 100 ms and 200 ms. These

subnets represent collaboration sites located in Europe, South America, or Asia. The third category consists of subnets that have 200ms plus RTTs. These subnets represent a few faraway collaboration sites located in Asia and Europe.

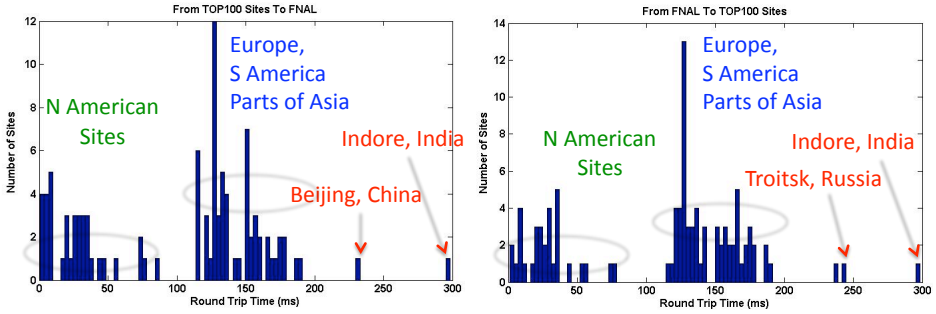


Fig. 2 RTT Statistics between FNAL and TOP 100 /24 subnets in Inbound and Outbound Directions.

RTT plays a key role in the TCP transfer throughput. Researchers [19][20] have calculated TCP throughput as a function of packet loss rate and RTT. In [19], a TCP throughput equation was derived for TCP Reno/New Reno:

$$Reno\_Thru = S/(RTT * \sqrt{2*b*p/3} + (T_0 * (3*\sqrt{3*b*p/8} * p * (1+32*p^2)))$$

where Reno_Thru is the transmit rate in bytes/second, S is the average packet size in bytes, RTT is the round trip time in seconds, p is the loss event rate (between 0 and 1.0, representing the number of loss events as a fraction of the number of packets transmitted), and T_0 is the TCP retransmission timeout value in seconds. We further simplify this by setting $T_0 = 4 * RTT$. The number of packets acknowledged by a single TCP ACK, which is usually two, is represented by b. This equation has been widely used in TFRC (TCP Friendly Rate Control) rate calculation.

Therefore, given the same network conditions, TCP throughput is inversely proportional to RTT. Usually, once the sender and the receiver are given, RTT is difficult to reduce unless a circuitous path is chosen. However, many factors in the real world (e.g., lack of peering, policy routing, and traffic engineering) can lead to circuitous paths and result in inflated RTTs. To give a better understanding of this problem, we ran traceroute from Fermilab to Indore, India, which has a RTT close to 300ms. Based on the traceroute results, we tagged the intermediate hops with geographic information obtained from a commercial geolocation database [21] and mapped them into Goggle Maps as shown in Figure 3. For simplicity, Figure 3 only demonstrates the path segments in US, which clearly shows a circuitous path from Fermilab to Indore, India.

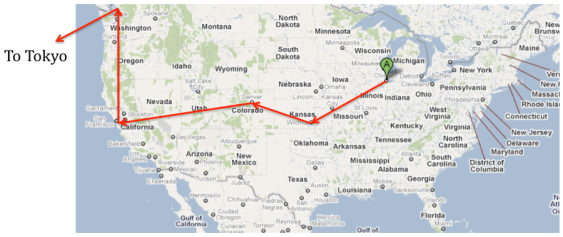


Fig. 3 Circuitous Path From Fermilab to Indore, India

To investigate the situation of circuitous path between Fermilab and its collaboration sites, we tagged the top 100 /24 subnets with geographic information obtained from a commercial geolocation database. We then calculated the straight distance between Fermilab and these subnets. Prior work [22] on the geolocation of Internet hosts observed that the expected RTT through the Internet is that obtained when bits travel the 4/9th the speed of light in vacuum. We assume that there was a direct path between Fermilab and its collaboration sites, and computed the corresponding RTTs. We referred to the RTT computed based on this assumption as Geo-RTT. We compared the computed Geo-RTT with the real RTT for each /24 subnet in both the inbound and outbound directions. If a subnet's real RTT is more than 1.5 times of the corresponding Geo-RTT, we would say there is a circuitous path

between Fermilab and the collaboration site. We coalesced the subnets that belong to the same collaboration sites. In the inbound direction, there are 25 collaboration sites that have circuitous paths to Fermilab. In the outbound direction, there are 21 such collaboration sites. Admittedly, it is difficult to completely eliminate the circuitous path situations due to geography constraints or other limitations. However, with careful selection of BGP peering points and/or routing policies, many circuitous path situations can be minimized.

3.3. Average Throughput Statistics

We calculated statistics for single flow's average throughput between Fermilab and its collaboration sites in inbound and outbound directions. Because each flow record includes data such as the number of packets and bytes in the flow, as well as the timestamps of the first and last packet, calculation of throughputs for an identified bulk data movement is simple. However, two additional factors must be considered. First, because a TCP connection is bidirectional, it will generate two flow records, one in each direction. In practice, a bulk data movement is usually unidirectional. Only the flow records in the forward direction record the true data transfer activities. The flow records in the other direction simply record pure ACKs of the reverse path, which should be excluded from throughput calculations. These flow records can be easily filtered out by calculating their average packet size, which is usually small. Second, a bulk data movement usually involves frequent administrative message exchanges between sites. A significant number of flow records are generated due to these activities. These flow records usually contain a small number of packets with short durations; the calculated throughputs are usually inaccurate and vary greatly. These flow records are also excluded from our throughput calculation.

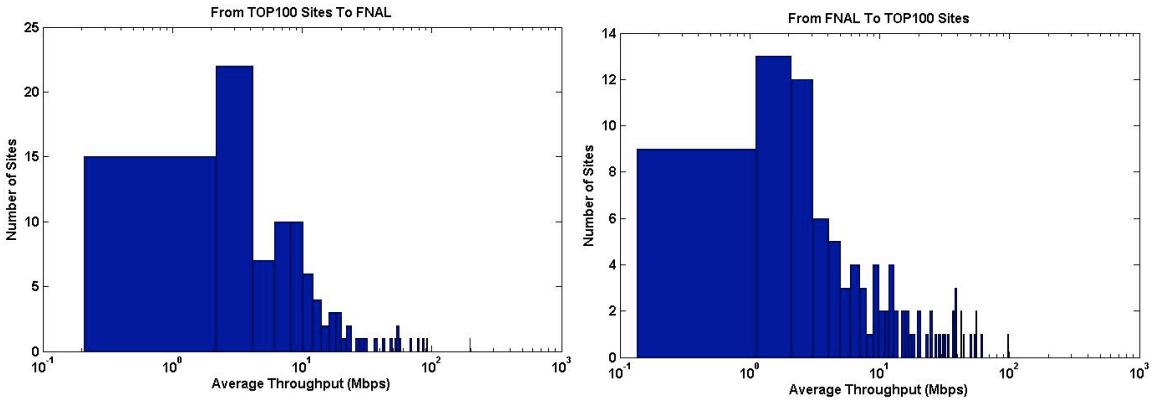


Figure 4 Ave. Thru. Statistics for Single Flow between FNAL and TOP100 Subnets

The average throughput statistics are shown in Figure 4. In the inbound direction, 2 collaboration sites' average single flow throughputs are less than 1Mbps. 43 collaboration sites' average throughputs are less than 10Mbps. Only 1 site's average throughput is greater than 100Mbps. In the outbound direction, 7 collaboration sites' average single flow throughputs are less than 1Mbps, 69 subnets' average throughput are less than 10Mbps. No site's average throughput is greater than 100Mbps. For those sites that have an average throughput of less than 1Mbps, we contacted their network administrators to conduct end-to-end performance analysis. Five sites responded to our requests. The end-to-end performance analysis indicated poor network conditions between these sites and Fermilab. To our surprise, one site in Greece is even connected to the outside world with a 100 Mbps link.

3.4. Parallel Data Transmission

Because TCP does not work well in LFPs (long fat pipes), data transmission achieves much greater use of bandwidth by allowing multiple simultaneous TCP streams. Parallel data transmission tools such as GridFTP have been widely applied to bulk data movement. On the other hand, bulk data movement usually involves multiple networked computer systems. Therefore, to monitor and analyze the status and patterns of bulk data movement between Fermilab and its collaboration sites, it is useful to understand the number of networked computer systems involved at each collaboration site generating the flow statistics. For this purpose, we binned traffic at 10-minute intervals and collected the flow statistics and the statistics of the number of different IPs involved at each /24 subnet in the inbound and outbound directions. The statistics are shown in Figure 5 and 6, respectively.

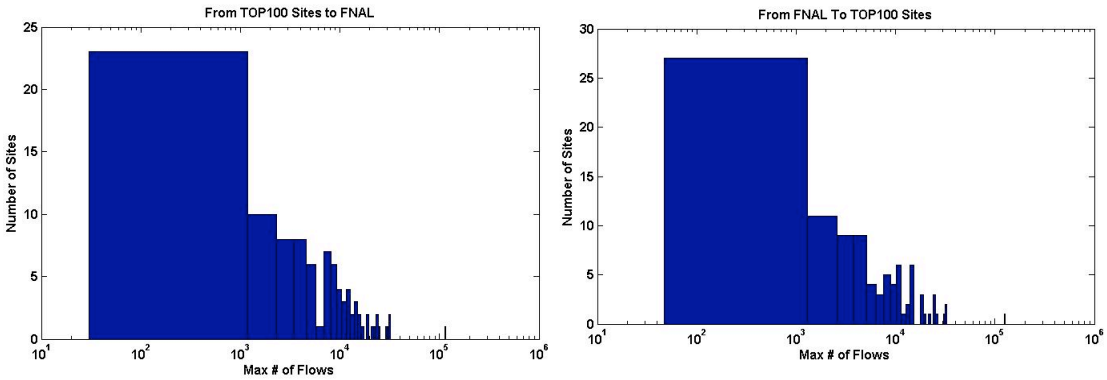
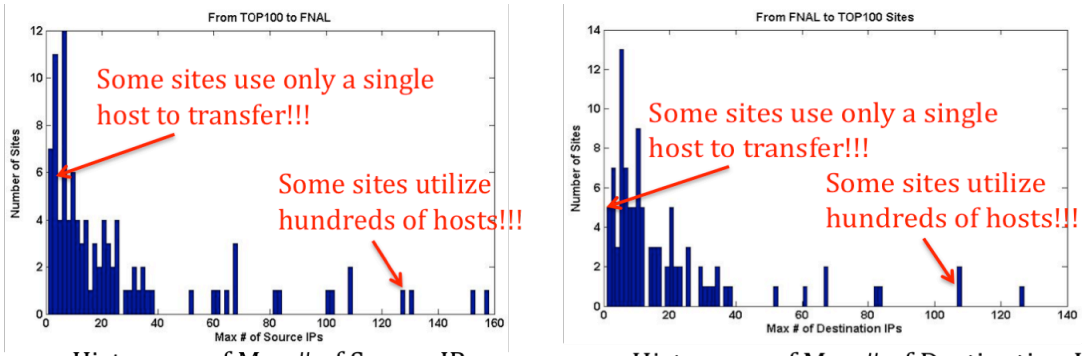


Fig. 5 Histogram of Maximum Number of Flows between Fermilab and Top 100 Off-site /24 Subnets



Histogram of Max # of Source IPs Histogram of Max # of Destination IPs

Fig. 6 Histogram of Maximum Number of IP addresses Involved at Top 100 Off-site /24 Subnets

It can be seen from Figure 5 and 6 that the bulk data movements across collaboration sites vary greatly in terms of scale and capability. For some large collaboration sites, the bulk data movements involve dozens, perhaps hundreds, of systems, and consist of hundreds, perhaps even thousands, of parallel data streams. On the other hand, small sites have very limited computing and networking resources. Some sites use only a single host with a small number parallel streams to transfer to/from Fermilab.

4. Conclusion

Fermilab is the US-CMS Tier-1 Centre and the main data centre for a few other large-scale research collaborations. Scientific traffic (e.g., CMS) dominates the traffic volumes in both inbound and outbound directions. Fermilab has deployed a Flow-based network traffic collection and analysis system to monitor and analyze bulk data movements between Fermilab and its collaboration sites. In this paper, we analyze the current status and patterns of bulk data movement between Fermilab and its collaboration sites.

5. Reference:

[1] OWAMP website, http://www.internet2.edu/performance/owamp/
[2] K. Papagiannaki, R. Cruz, C. Diot, "network performance monitoring at small time scales," Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, Miami Beach, FL, USA, 2003.
[3] T. Benson, A. Anand, A. Akella, M. Zhang, "Understanding Data Center Traffic Characteristics," In Proceedings of ACM WREN, 2009.
[4] V. Paxson, "Automated packet trace analysis of TCP implementations," In Proceedings of SIGCOMM'97, 1997.
[5] V. Paxson, "End-to-End Internet packet dynamics," In proceedings of SIGCOMM'97, 1997.
[6] NetFlow website, http://www.cisco.com/
[7] B. Choi, S. Bhattacharyya, "On the accuracy and overhead of Cisco Sampled NetFlow," In Sigmetrics Workshop on Large Scale Network Inference (LSNI): Methods, Validation, and Applications (June 2005).
[8] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural Analysis of Network Traffic Flows. In ACM SIGMETRICS'04, New York, June 2004.
[9] A. Kalafut, J. Merwe, M. Gupta, "Communities of interest for internet traffic prioritization," In Proceedings of 28th IEEE International Conference on Computer Communications Workshops, 2009.
[10] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing Network-Wide Traffic Anomalies," In Proceedings of ACM SIGCOMM'04, 2004.
[11] A. Lakhina, M. Crovella, C. Diot, "Characterization of Network-Wide Anomalies in Traffic Flows," In Proceedings of IMC'04, 2004.
[12] J. Wallerich, H. Dreger, A. Feldmann, B. Krishnamurthy, W. Willinger. A methodology for studying persistency aspects of internet flows," SIGCOMM Comput. Commun. Rev. 35, 2 (2005).
[13] Internet2 NetFlow, http://netflow.internet2.edu/weekly/
[14] R. Sommer and A. Feldmann, "NetFlow: Information loss or win?," in Proc. ACM Internet Measurement Workshop, 2002.
[15] C. Gates, M. Collins, M. Duggan, A. Kompanek, M. Thomas, "More netflow tools: for performance and security," In Proceedings of LISA'04, 2004.
[16] V. Krmicek, J. Vykopal, R. Krejci, "Netflow based system for NAT detection," In Proceedings of 5th international student workshop on emerging networking experiments and technologies, 2009.
[17] PhEDEx – CMS Data Trnasfers, http://cmsweb.cern.ch/phedex/prod/.
[18] SiLK Toolset website, http://tools.netsa.cert.org/silk/silk.html.
[19] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, "Modeling TCP Throughput: A Simple Model and its Empirical Validation," Computer Communication Review, Vol. 28, Num. 4, 1998.
[20] M. Mathis, J. Semke, J. Mahdavi, T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," ACM SIGCOMM Computer Communication Review, Vol. 27, Issu. 3, 1997.
[21] www.maxmind.com
[22] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP geolocation using delay and topology measurements. In IMC, 2006.
[22] A.Bobyshev, M.Grigoriev, Methodologies and techniques for analysis network flow data, CHEP04, CERN, Interlaken, Switzerland, 27th September – 1st October 2004
[23] A.Bobyshev, D.Lamore, P.Demar, A real-time system for detecting and blocking of network scanning based on analysis of netflow data, CHEP04, CERN, Interlaken, Switcherland, 27th – September – 1st October, 2004
[24] A.Bobyshev, P.Demar, V.Grigaliunas, M.Grigoriev, Use of Flow data for Traffic Analysis and Network performance characterization, CHEP07, Victoria BC, Canada, 2-7 September 2007

[25] A.Bobyshev, P.Demar, W.Wu, Real Time Flow Analysis for Network Services, CHEP09, 21-27 March 2009, Prague, Czech Republic